日中韓辭典研究所
The CJK Dictionary Institute

DiaLEX:

ARABIC DIALECTS FULL-FORM LEXICON

Overview

While Modern Standard Arabic is used as the official language of 22 Arab nations, Arabs normally use one of the 30 or so modern dialects for communicating with family and friends in their daily life. *DiaLEX* is the most comprehensive Arabic computational lexicon ever created for Arabic dialects. Designed for NLP applications like MT, NER and morphological analysis, it is ideally suited for training speech technology models.

Coverage

DiaLEX covers or will eventually cover the following major Arabic dialects: Egyptian, Kuwaiti, Qatari, Emirati, Saudi Arabian Najdi, Saudi Arabian Hejazi, and Palestinian.
It is rich in morphological, grammatical, phonological, and orthographic attributes (currently about 100 million for three dialects). In addition, it maps all unvocalized forms to their vocalized counterparts and to the lemma, and provides phonemic transcriptions and graphemic transliterations.

Dialect	Lenmata	Entries
Egyptian	25,000	35 million
Hejazi	25,000	35 million
Emirati	25,000	35 million

Distinctive Features

- Extremely comprehensive some 40 million full form entries entries per dialect
- Rich in morphological attributes: all inflected and cliticized and negated forms
- Numerous orthographic variants
- Includes high frequency proper nouns (personal names and place names)
- Fully vocalized and unvocalized Arabic
- Accurate phonemic phonetic transcriptions and transliteration
- All wordforms are cross-referenced to their lemma

日中韓辭典研究所
The CJK Dictionary Institute

ARAB_V	ARAB_BW	LEMMA_V	POS	GEN	NUM	NPG
بِيتْ	biyto	بِيتْ	N	M	S	000
اِلْبِيتْ	Ailobiyto	بِيتْ	N	M	S	000
بِيتِي	biytiy	بِيتْ	N	M	S	S1C
بِيتَكْ	biytako	بِيتْ	N	M	S	S2M
بِيتِكْ	biytiko	بِيتْ	N	M	S	S2F
بِيتُو	biytuw	بِيتْ	N	M	S	S3M
بِيتْهَا	biytohaA	بِيتْ	N	M	S	S3F
بِيتْنَا	biytonaA	بِيتْ	N	M	S	P1C
بِيتْكُو	biytokuw	بِيتْ	N	M	S	P2C
بِيتْهُمْ	biytohumo	بِيتْ	N	M	S	P3C

The CJK Dictionary Institute

The CJK Dictionary Institute (CJKI) was founded in 1993. Its principal activity is the compilation of large-scale dictionary databases of proper nouns and technical terms for CJK (Chinese, Japanese, Korean) and Arabic, currently with over 50 million entries. CJKI has become the world's prime source for CJK lexical resources for the IT industry and software developers, providing high-quality comprehensive dictionary data, educational tools, and consulting services. Based in Saitama, Japan, CJKI is headed by Jack Halpern, editor in chief of *The Kodansha Kanji Learner's Dictionary* and several other dictionaries that have become standard works for learning Japanese.

Jack Halpern (春遍雀來), CEO of The CJK Dictionary Institute, is a lexicographer by profession, specializing in Japanese and Chinese. His work as an editor in chief of learner's dictionaries resulted in various renowned standard reference works. He has been a resident of Japan for over 40 years, but was born in Germany and has lived in France, Brazil, Japan, and the United States. He is an avid polyglot who has studied 18 languages (speaks 11).

The CJK Dictionary Institute, Inc.
34-14, 2-chome, Tohoku, Niiza-shi, Saitama 352-0001 JAPAN
Phone: 048-473-3508 Fax: 048-486-5032
Email: jack@cjki.org Web: www.cjk.org